Data Explosion, Data Nature and Dataology

Yangyong Zhu[1], Ning Zhong[2], and Yun Xiong[1]

[1] School of Computer Science, Fudan University
Shanghai 200433, P.R. China
yyzhu@fudan.edu.cn, yunx@fudan.edu.cn

[2] Dept. of Life Science and Informatics, Maebashi Institute of Technology
Maebashi-City 371-0816, Japan
zhong@maebashi-it.ac.jp

Abstract. The essence of computer applications is to store things in the real world into computer systems in the form of data, i.e., it is a process of producing data. Some data are the records related to culture and society, and others are the descriptions of phenomena of universe and life. The large scale of data is rapidly generated and stored in computer systems, which is called *data explosion*. *Data explosion* forms *data nature* in computer systems. To explore data nature, new theories and methods are required. In this paper, we present the concept of data nature and introduce the problems arising from data nature, and then we define a new discipline named *dataology* (also called *data science* or *science of data*), which is an umbrella of theories, methods and technologies for studying data nature. The research issues and framework of *dataology* are proposed.

1 Introduction

According to the recent IDC research report entitled "As the Economy Contracts, the Digital Universe Expands" [1], the amount of new digital information reached about 486 billion gigabytes in 2008 and increased 3 percent faster than IDC previous projection. The digital universe is expected to be double in size every 18 months. In 2012, five times as much digital information will be generated versus 2008. When the data are explosively increasing, they also become more complicated and diversified. At the *IBM Information on Demand 2009* conference, experts pointed out that in the world almost 15 GB (gigabytes) data are produced every day. These data come from various equipments including sensors, RFIDs, meters, GPSs, etc., and at least 80 percent of new data are unstructured, such as Web contents, Web logs, email, image, video, audio, and so on.

All the facts mentioned above indicate that data explosion has happened and been spreading. In fact, *data explosion* is the course that data in computer systems explosively increase since human continuously stores data when using the computers. During the course of data explosion, the mass data appear multiple natural features including out of control, unknown, diversity and complexity. Therefore, data explosion forms *data nature*. Studying data nature is an effective

N. Zhong et al. (Eds.): BI 2009, LNAI 5819, pp. 147–158, 2009.
© Springer-Verlag Berlin Heidelberg 2009

approach to study real nature, for example, the researches on *Bioinformatics* [2], *Brain Informatics* [3] and *Behavior Informatics* [4] indicate that we can study life and human brain by studying related scientific data.

This paper proposes *dataology* (also called *data science* or *science of data*) which is an umbrella of theories, methods and technologies for studying data nature. The rest of this paper is organized as follows. Section 2 introduces data explosion. Section 3 describes data nature including natural features and evolution, as well as key issues in data nature. Section 4 presents dataology as a new research discipline and provides its framework and content. Finally, Section 5 gives concluding remarks.

2 Data Explosion

Data are increasing explosively with the development of human being. Trying to remember is the instinct of human being. From time immemorial, using brain to remember experienced things is the primary means. Because of some unknown reasons, human memory cannot retain everything they read. The memories in human brain are also unreliable. Thus, human seeks various tools to help them to memorize all along. Originally, human carved figures and characters on hard objects to assist remembering. They found that the information recorded out of brain was convenient for transmission and communication, therefore the human instinct of recording information is deepened.

The inventions of papermaking and printing brought about the first data explosion [1] during which a mass of natural things (including natural phenomena, culture, society, etc.) are represented by characters or figures, and then printed into books or materials. That is, the information about a period of history can be recorded into a book for memorizing and transmitting, such as the Bible, the Records of the Grand Historian, and so on. The information can be stored for a long time, replicated many times, and spread widely. During the course of this data explosion, the authors/publishers produced information, and the books/libraries stored and conveyed information.

The inventions of computers (especially Internet/WWW) and storage devices brought about the second data explosion. It is a process that data in computer systems explosively increase because human continuously stores data when they use the computers. During this explosion, all of books in the earlier libraries and previous publications (i.e., main productions in the first data explosion) can be stored into a personal computer, even a removable hard disk.

So far no proof indicates that there exists a kind of device which can replace computers and storage devices. In the future, a certain kind of man-made being

[1] The term "information explosion" and "data explosion" are usually replaceable. Information is the explanation of data, and data is the symbolic representation of information. Therefore, in general, the more the information, the more the data required to be stored is. Conversely, the more the data, the more the information can be expressed is. In this paper, we use the term "data explosion".

would be created, who has the camera memory and its memory retains everything it reads. It would exhibit powerful processing ability. This will bring the next data explosion.

Data explosion makes people lost in the mass data, which can be illustrated in the following aspects:

– Difficult to ensure the truth of data: we cannot know if the data obtained from computer systems (such as Internet/WWW) are true. This will lead that we do not know which are usable though the mass data we own.
– Difficult to share data: data sharing becomes more and more difficult though it is one of goals of computer systems to provide the ability of data sharing. A great amount of data are produced every day, we do not know what need to be shared, as well as how to share.
– Difficult to keep data consistent: we cannot ensure the consistency of data. For example, we often get the different results when we query the same object on different websites.

Though we are continuously developing new technologies, such as grid computing [5], cloud computing [6]), and wisdom-Web computing [7], the above problems become more and more serious.

3 Data Nature

3.1 What Is Data Nature?

The development of the world is the course in which human being continuously explores real nature (universe and life) and builds up culture and society. When the human activities and their results are stored into computer systems, we create a data nature unconsciously.

During data explosion, more and more data are stored into computer systems. They have various categories and formats. For example, there exist the following data categories:

– Personal data: The personal data are often stored in personal computers, including personal privacy data and work data. The personal work data involve various contents, such as the the data for the company/organization, the data for the job, and so on. The personal data are also dispersed in the Internet, which are often neglected though it is a kind of important personal data.
– Enterprise data: The enterprise data are stored in enterprise computer systems. They are respectively from enterprise operations, clients, competitors, trades, etc.
– Government data: The government data are stored in governmental computer systems, which include government operation data, social resource data, economics of population data, etc.
– Public data: The public data are stored in public websites, which can be accessed by Web search engines.

– Data in various kinds of languages: Different nationalities use different languages, such as English, Chinese, Arabic, and so on. Therefore, the data in different languages have been produced.
– Geographical data: The geographical data describe the geographical situations and changes in our countries or areas, which include the data related to space, ocean, earth, country, etc.
– Life data: The life data implicate a great deal of life characteristics and information, which include DNA sequence data, protein sequence data, and cognitive and medical brain data, etc.
– Cultural and social data: The cultural and social data describe the development of the human being and the society, which include human behaviors, economic data, etc.
– Internet data: The Internet data are dispersed in the Internet. They are easy to access, but contain numerous garbage and viruses. The appearance of Internet data makes the data in computer systems to show more natural features.

The formats of the data in data nature include:

– Special format data: The special format data are the data produced by professional digital equipments, such as medical image data (including x-ray, MRI, CT, EEG, etc.), cognitive brain image data, GIS data, multimedia data, which can be collected and processed by professional equipments or software.
– General format data: The general format data are the data stored in general format. In the early stage of computer applications, most of data are stored in relational databases and managed by general database management systems, such as Oracle or DB2. These data have clear structures and are easy to process.

The mass unknown data in computer systems are the basis of data nature. We do not know if the data from the Internet are true; we query the same object on different websites, but often get different results; maybe in the Internet a database has indicated that human being will face energy crisis, but we cannot grasp this knowledge; we input DNA sequences into computers, but we do not know what they indicate? what law they have? which fragments of DNA make the differences among people? how genes change during the evolution of species? whether gene evolution or mutation exists? and so on.

Though human being have produced data and are continuously producing data, these data have shown various natural features as follows:

– Out of control: The explosive increase of data leads that human cannot completely control it. Human also cannot control the appearance and spread of computer viruses, the deluge of SPAMs, the jam of NII (National Information Infrastructure) [8] due to network attack, etc.
– Unknown: Since HGP (Human Genome Project) [9] launched, a great amount of DNA data are stored into computers. However, people do not know what

these DNA sequences indicate, as well as how genes change during the evolution.
– Diversity: As mentioned above, there are various data categories, they are from space, ocean, biology, brain, multiple languages, various trades, as well as in the Internet/out of the Internet, public/private, therefore, the data in data nature exhibit diversity.
– Complexity: The mass data in computer systems are complex, they have various data formats and there exist many associations and relationships among these data.

Thus, natural features of data are more and more obvious. Various data forms and data “regions/areas” or data “tribes” have come into being. Though during data explosion we cannot distinctly describe the data nature, in fact, we have already worked and lived in it. The Digital Earth project [10] is in process. Through it, we are gradually transforming real nature into data nature.

3.2 Key Issues in Data Nature

As shown in Fig. 1, our culture and society are built on real nature at the beginning, and then the computer science and technology help people store both “culture and society” and “real nature” into computer systems when the computer was invented (as shown in Fig. 2). This meets both practical requirements and the human instinct of remembering.

As shown in Fig. 3, culture and society will be built on both real nature and data nature, and supported by computer science and technology. It means that our culture and society will rely increasingly on data nature.

In data nature, we will face many new problems. For example, if one asks how many papers are related to DNA in *science*, it is easy to answer; if he asks how

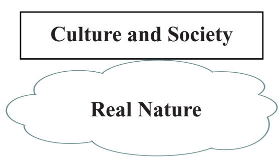

Fig. 1. Culture and society are built on real nature

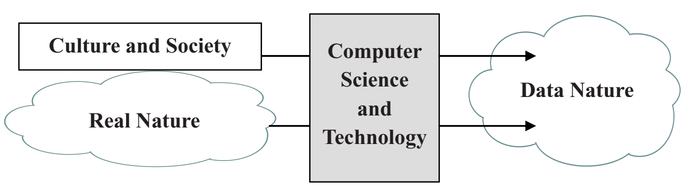

Fig. 2. From real nature, culture and society to data nature

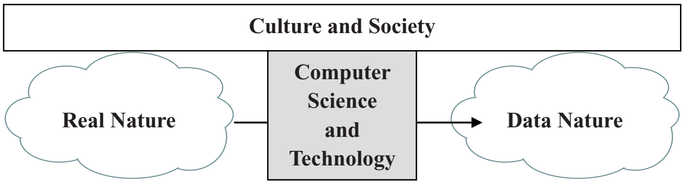

Fig. 3. The culture and society based on both real nature and data nature

many fields are studied in *science*, it is difficult to answer; if he asks which paper is associated with each other in *science* or what is indicated by all the papers in *science*, it is more difficult to answer. Therefore, we say that *science* itself is a valuable research topic. *Science* has been digitized and stored into computer systems in the form of data. It forms a "data tribe" or "region" like human tribe or region in real nature. Similarly, a library, a DNA database and a DOD (Department of Defense) system are bigger data tribes or regions which need us to explore and study.

In fact, we will face the following problems in data nature:

- How to recognize data nature?
- How to recognize real nature through data nature?
- How to live in both data nature and real nature?
- How to recognize whether the contents of data nature truly represent the real nature?
- How to develop and utilize data nature for the part superior to real nature?

Human existing foundation is material, that is, all the basic necessities of life, such as clothing, food, shelter and transportation, are based on materials. Computer systems have changed the foundation, for example, life habit, thinking, standpoint, morality, law, etc., have been changed. People rely increasingly on data in computers rather than the materials. Now human lives have been occupied by more and more data.

4 Dataology

4.1 What Is Dataology?

As mentioned above, the research achievements on real nature are stored into computer systems in the form of data which forms data nature. The exploration on data nature will be on a higher level than before. Many principles and laws in real nature, such as *prime number*, *fibanacci sequence*, *golden ratio*, *pareto principle*, etc, are also available in data nature (e.g., DNA data tribe). Many nice things in the world have been shown to us in the form of data, so now it is necessary for us to discover them from data.

With the forming of data nature, the science and technology of exploring data become more and more important. One significant discipline called *dataology* is forming, which takes data as a research object.

For a long time, the researches and technologies aiming at data mainly focus on data storage and management, such as database technology. Their primary goal is to model real nature, rather than to take data themselves as a research object.

The appearance of data mining technology [11] means that people began to study the laws aiming at data in computer systems. In the field of Internet, more and more researches focus on network behavior, network community, network search, and network culture. Because of the accumulation of data, newly disciplines, such as bioinformatics and brain informatics, are also typical dataology centric research areas. For instance, DNA data in bioinformatics are the data that describe natural structures of life, based on which we can study life using computers.

Brain informatics, which emphasizes a *systematic* approach to investigating human information processing mechanisms, including measuring, collecting, modeling, transforming, managing, mining, interpreting, and explaining multiple forms of brain data obtained from various cognitive experiments by using powerful equipments, such as fMRI (functional magnetic resonance imaging) and EEG (electroencephalogram) [3]. A tangible goal of brain informatics studies is to form a brain data area with a conceptual brain data model namely *Data-Brain*, which represents functional relationships among multiple human brain data sources, with respect to all major aspects and capabilities of human information processing systems (HIPS), for systematic investigation and understanding of human intelligence. In fact, Data-Brain construction is to form brain data area in computer. On one hand, developing such a Data-Brain is a core research issue of brain informatics. Systematic brain informatics research needs a Data-Brain to describe related knowledge and to annotate various human brain data sources, in order to support data sharing and data integration. Based on this way, it provides a long-term, holistic vision to uncover the principles and mechanisms of underlying HIPS. On the other hand, the brain informatics methodology supports such a Data-Brain construction. The Data-Brain is used to model various human brain data related knowledge, involving data themselves, data production and data disposal.

Thus, we need a special discipline *dataology* and pay much attention to studying data themselves. Dataology is an umbrella of theories, methods and technologies for studying data nature. It will provide basic theories and methods for many (maybe all) disciplines and fields, including *data acquisition*, *data analysis*, *data awareness*, *data management*, etc. These basic theories and methods will be applied on various fields for developing special theories, methods and technologies, which will form dataology for specific domains.

The main research issues of dataology are:

- To form data areas in various domains;

  Massive datasets have driven research, applications, and tool development in business, science, government, and academia. The continued growth in data collection in all of these areas ensures the fundamental problem of dataology addresses, namely how the component (i.e., data area) of data nature is formed in various domains. We are experiencing a strong demand for more powerful, intelligent computing paradigms for large-scale data measuring, collecting, modeling, transforming, exploring and managing [7].

- To study the structures of datasets in data nature;
  In data nature, a dataset which we will deal with may be a hard disk full of data, or a database managed by a certain DBMS, or a Web server, therefore, we need to study how to access these data from these media or circumstances. It is critical to analyze storage structure and logical structure of a dataset, which is called the study on the structures of a dataset.
- To acquire available data from data nature;
  Like to explore the golden or the oil, we will acquire available data from data nature. However, it is different from data mining because it only gets original data rather than process and analyze data. In general, available data should be acquired from various data sources in data nature and these data should be integrated. Sometimes they should be stored in the data warehouse after data cleaning.
- To prove the rules of data nature by theoretical methods;
  Like the research on science, we need to establish many theories and methods, and present hypothesis, induction, deduction and inference, and build up logical and theoretical systems in order to solve various problems generating from data nature.
- To discover the rules of data nature by experimental means;
  There are a lot of propositions and rules to be verified. Furthermore, many valuable results will produce through experiments, which is similar to the chemical experiments. Therefore, we need to establish various experimental systems and means in order to discover the rules in data nature.
- To develop and utilize data resources in data nature.
  Developing and utilizing data resources is a goal to research data nature, which will support the research on natural science and social science and serve for human life and social development. We believe that data resources are the most important resources in this century (perhaps more important than oil and coal), therefore, it is an important issue to develop and utilize data resources in data nature, which is an important topic in dataology.

4.2 The Framework of Dataology

The framework of dataology is shown in Fig. 4. This framework includes two main parts: foundations of dataology and applications of dataology (e.g., universal dataology, life dataology, behavior dataology, etc).

Foundations of Dataology. Foundations of dataology is composed of three aspects: data acquisition, data analysis and data awareness. They can be divided in more detail (see Fig. 4). All these technologies require data management. There are some existing technologies including data integration, data management (e.g., file system, database management system and data warehouse, etc.), data mining, data visualization, etc. They are developing continuously.

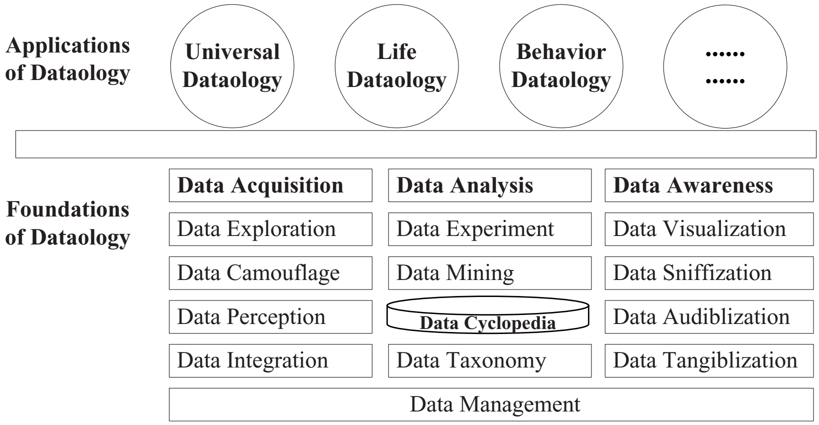

Fig. 4. The framework of dataology

In addition, dataology needs to develop more new technologies:

- **Data Experiment**
  Most of people consider that bioinformatics makes biological experiment to become data computing. However, we believe that bioinformatics is more like biological data experiment. For example, we have a gene sequence of SARS virus in which "A", "C", "G" and "T" are represented as the points with different colors. When we optionally change sequences or duplicate parts of sequences (this just like to try to mix multiple reagents in chemical experiments), we may find an "S" picture. If that is quite true, the result is very valuable. Such cases also exist in other fields (e.g., brain informatics). This means that data experiment technology will be on demand.

  Data experiment is to use various known or unknown methods to deal with a dataset in order to discover special features and laws. It focuses on the randomness of methods and the unpredictability of results. This is different from data mining in which the selection of methods is based on the prospective results.
- **Data Camouflage and Perception**
  The data camouflage is to camouflage the private data which are exposed in the public. Different from data security (privacy protection, privacy mining), data camouflage focuses it efforts on camouflaging the data in the public, rather than storing data in a safe place to prevent invasion.

  In the field of computer, a logion says, "garbage in garbage out". However, current problems are that we do not know which data are garbages, as well as which data are the disguise of valuable data. How to obtain the valuable data under this circumstance? This involves data perception which is to percept the camouflaged data. It can be regarded as the reverse of data camouflage.

- **Data Taxonomy**
  Data taxonomy is to classify data to form the pedigree of data and history of development. Because of the forming of data nature, the data exhibit diversity and have various categories. The data in data nature need to be classified according to data purpose or relationships, which is similar to classify the things in real nature according to species, history or culture. This new technology is called data taxonomy.
- **Data Awareness**
  The sense of human being includes vision, audition, sniff, touch, etc. Thus, if people want to feel data nature as feeling real nature, only data visualization technology is not enough. Dataology needs to develop new technologies of data awareness including data sniffization, data audibization, data tangiblization, etc.

Applications of Dataology. Because of the forming of data nature, people are unable and unnecessary to use all of data and pursue data consistency. People will use data in part under their lives and work circumstance. Thus, for obtaining the better results, new data technologies will be the special technologies aiming at different fields and circumstances instead of the general technologies. For example, the technologies of bioinformatics are just the special technologies. Therefore, specific domain dataology will create corresponding data technologies.

The applications of dataology consist of universal dataology, life dataology, behavior dataology and so on, in which, some specific domain dataologies have been formed from the viewpoint of dataology, because they all work on data nature. For example, we believe that both bioinformatics and brain informatics should belong to life dataology, and behavior informatics should belong to behavior dataology. In addition, various dataologies will come into being, such as spatial dataology, oceanic dataology, financial dataology, and so on.

Let us take *brain informatics* also called *brain dataology* as an example to illustrate the applications of dataology in the framework. From the viewpoint of dataology, the key steps in brain dataology research are shown in the followings:

1. Brain activities experiments:
   To do the experiments aiming at brain activities using some devices, such as fMRI, OT, ERP/EEG, etc, and collect the experimental data which will be stored into the data nature in the computer systems.
2. Brain data collection:
   To collect the brain data from data nature, for example, we can collect the brain data through the experiments, the literatures or the public databases.
3. Brain data integration:
   To integrate various types of brain data and build the brain data warehouses according to the demands on the brain science research in order to provide the comprehensive data for systematical research.
4. Brain data mining:
   To discover the rules in human brain activities when people participate the cognition activities, such as, solving problems, reasoning, making decisions,

learning, etc. In general, new data mining algorithms should be developed according to the demands on brain science research.

5. Mining results analysis:
   To analyze the brain activities rules and cognition approaches from the results of data mining, because the mining results indicate the rules in the brain activity data. In this stage, brain scientists, psychologists and dataologists are expected to collaborate.
6. Mining results verification:
   To do the experiments and verify the results from brain data mining (i.e., the potential rules of brain activities). If it is true, the knowledge of brain activities will be accepted, conversely, the brain activities experiments in Step 1 should be improved, for example, designing the newly instruments and devices.
7. Brain activities experiments improvement:
   To improve the brain activities experiments according to the verified results in Step 6. And so on, over and over again, human recognition ability on the brain is improved gradually.

In the above steps, the tasks in Steps 1, 6 and 7 are administrated by brain scientists, or brain scientists are expected to participate, whereas, the tasks in Steps 2, 3, 4 and 5 are performed by dataologists.

It can be expected that all of the existing fields can form the corresponding dataology. Thus, we say that dataology is one of the most important disciplines in the 21^{st} century.

5 Conclusion

We explore real nature and use computers to represent our discoveries, society, nature and human being. Data have been produced explosively and a complex data nature has been created unconsciously. The history of human being and society will become the history of data. Therefore, it is necessary for human to research data nature, and new theories and methods are required, which is the goal to present the new discipline "*dataology*".

Data nature (i.e., data in computer systems) is the research object of *dataology*. There will be many new research contents and methods, such as data experiment, data taxonomy, data awareness, and so on. In the 21^{st} century, people will live in both real nature and data nature. Dataology as an emerging discipline will become one of the most important disciplines. The data technology also will become one of the most important technologies in the future.

Acknowledgement

The research was supported in part by Shanghai Leading Academic Discipline Project under Grant No. B114. The authors would like to thank professor Yixue Li, professor Yang Zhong and professor Longbing Cao for their suggestions. And we would like to thank Jianhui Chen and Xue Bai for collecting materials.

References

1. http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm (accessed May 2009)
2. Krawetz, S., Misener, S. (eds.): Bioinformatics Methods and Protocols. Humana Press (2000)
3. Zhong, N., Liu, J., Yao, Y.Y., Wu, J., Lu, S., Li, K. (eds.): Web Intelligence Meets Brain Informatics. LNCS (LNAI), vol. 4845, pp. 1–31. Springer, Heidelberg (2007)
4. Cao, L.B.: Behavior Informatics and Analytics: Let Behavior Talk. In: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops (2008)
5. Berman, F., Fox, G., Hey, A. (eds.): Grid Computing: Making the Global Infrastructure a Reality. John Wiley & Sons, Chichester (2003)
6. Hayes, B.: Cloud Computing. Communications of the ACM 51(7), 9–11 (2008)
7. Zhong, N., Liu, J., Yao, Y.Y.: Envisioning Intelligent Information Technologies through the Prism of Web Intelligence. Communications of the ACM 50(3), 89–94 (2007)
8. Chapman, G., Marc, R.: The National Information Infrastructure: A Public Interest Opportunity. Computer Professionals for Social Responsibility 11(2), 13–15 (1993)
9. Collins, F.S., Patrinos, A., Jordan, E., et al.: New Goals for the U.S. Human Genome Project: 1998-2003. Science 282(5389), 682–689 (1998)
10. Freeston, M.: The Alexandria Digital Library and the Alexandria Digital Earth Prototype. In: Proceedings of the 4th ACM/IEEE-CS joint Conference on Digital Libraries (2004)
11. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: an overview. In: Advances in Knowledge Discovery and Data Mining (1996)